Detailed Annotation Guideline

Brief Annotation Plan (taken from Project Proposal)

Given our limited time and team size, we will focus on a single high-value annotation strategy. We have below three approaches that we have considered, from which we will select one based on feasibility and impact (we've bolded what we believe is the most feasible for this project):

- *Document-Level Factuality & Misleading Language*: Instead of verifying each claim individually, we would assess the overall credibility of an article using a factuality rating scale (1 = would be completely false, 5 = fully factual). We could also mark whether the article uses misleading tactics such as exaggeration, emotional manipulation, lack of sources, or logical fallacies, although this may be time-consuming and difficult to operationalize.
- *Source Attribution & Citation Behavio*r: alternative, we plan on annotating whether news articles include citations, whether those citations are real, and whether they reference authoritative sources. This approach will help assess the reliability of AI-generated news, and is quite feasible at a document-level.
- ***Persuasive Framing & Emotional Appeal*: We could alternatively categorize the dominant rhetorical techniques used in each article/piece, identifying whether the primary persuasive strategy is logical argumentation (logos), emotional appeal (pathos), authority-based persuasion (ethos), or sensationalism. This would allow us to compare how AI and human writers may structure persuasive narratives differently, and also allows us to use this type of annotation across news articles and opinion pieces alike.**

Annotation Guideline:

We will perform annotations on Persuasive Framing and Emotional Appeal. Given a body of text (length should less than 60 words), we will ask the annotator to label the text categorized as one of:

- Logical Argument (Logos)
- Emotional Appeal (Pathos)
- Authority-Based Persuasion (Ethos)

We will use a combination of expert annotations (i.e. annotations done by us) and annotations done by MTurk workers on Cloud Research. If necessary, we are also considering utilizing LLM to automate part of the annotation process, providing the model with few shot prompting to ensure high quality, accurate annotation. Moreover, we will also conduct 2 manual evaluation on the annotated data:

1. For MTurk worker-labelled annotations, we (the experts) will randomly sample part of the annotated data and compare against results we obtain. If the inter-annotator agreement score (e.g. Cohen's Kappa) is high, we can verify with confidence that the annotation done by the worker is of high quality.
2. To verify the quality of the annotation done by GPT, we will also conduct random sampling, perform annotations, and calculate inter-annotator agreement score and precision/recall and F1-scores. If the inter-annotator agreement is high and the F1-scores are also reasonable, we can assume our LLM model performed reasonably well at annotating.

The annotators should be from an English-speaking nation, and speak English as their primary language. They should also have an approval rating of 90 or higher and have performed at least over 1000 annotation tasks on the platform.

Example:

Say for example, we have 1000 items to annotate (HITs - Human Intelligence Tasks).

1. We will split our annotation task into 5 batches, each part having 200 unique items to annotate. We (experts) will manually annotate 10% (i.e. 20 HITs) and calculate the inter annotator agreement for each of the 5 batches.
2. If the agreement score is high, we will release another 10% for each batch to be annotated by MTurk workers (there should be 3 workers who annotate the same task for majority voting), annotate it also ourselves, and calculate the annotator agreement between the workers and ourselves, as well as among the 3 workers.
3. If the agreement score is high, we will use GPT-4o to annotate the rest, sample 10% and do another calculation of agreement scores and also calculate Precision/Recall and F1 scores.
4. If time permits, we can use a different LLM model to annotate the same HITs annotated by GPT-4o and compare results similarly.